Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2016 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 6-27-2016

STOCK PREDICTION VIA SENTIMENT AND ONLINE SOCIAL STATUS

Jih-Shin Chen
National Taiwan University, d02725003@ntu.edu.tw

Hsin-Min Lu
National Taiwan University, luim@ntu.edu.tw

STOCK PREDICTION VIA SENTIMENT AND ONLINE SOCIAL STATUS

Jih-Shin Chen, Department of Information Management, National Taiwan University, Taipei, Taiwan, d02725003@ntu.edu.tw

Hsin-Min Lu, Department of Information Management, National Taiwan University, Taipei, Taiwan, luim@ntu.edu.tw

Abstract

Studies of stock market prediction show that stock movements are related to the sentiment of social media. However, few studies have investigated the role of online social relations in predicting stock movements. This paper aims at constructing features that capture users' online social status and incorporating these into stock prediction models. Online opinions are often developed through interactions and are weaker in their early stages. We developed a feature-enhancing procedure motivated by statistical surveillance approaches to strengthen the ability to capture emerging trends. We evaluated our feature-enhancing procedure by developing models to predict stock returns in the following 20-minute period. A comparison of experimental results with baseline models shows that our feature-enhancing design helped to predict stock movements. The model (SE_CUSUM) that adopted features enhanced by cumulative sum (CUSUM), a statistical surveillance approach, performed better than baseline models in terms of directional accuracy, balanced error rate, root mean square error, and mean absolute error. Our simulated trading also showed that SE_CUSUM realized a higher profit than the baseline approaches. These results suggest that incorporating online social status and our feature-enhancing procedure improve high frequency stock prediction performance.

Keywords: Social relations, Social media, Stock predictions, CUSUM, EWMA.

1. INTRODUCTION

This study investigates the significance of online social relations in predicting stock market dynamics. We develop features that consider both the sentiment of a message as well as the online social status of the poster. We process these features using statistical surveillance approaches, such as the cumulative sum (CUSUM), to boost weak signals embedded in them. The processed features are then adopted in subsequent prediction and simulated trading experiments to study their utility.

Recent studies have documented the relationship between social media and stock markets (Antweiler & Frank, 2004; Bollen et al., 2011; Das & Chen, 2007; Q. Li et al., 2014; Smailović et al., 2014). These studies have investigated how sentiment, message volume, and other text features influence stock markets. However, less attention has been paid to the role of online social relations in stock predictions.

Our study identifies two issues in adopting online social relations and sentiment features from social media. First, the opinions from users with higher online social status may be more influential to the community and thus more important for predicting stock dynamics. Second, online opinions are often developed through interactions and are weaker in their early stages. Directly adopting features that capture online opinion may be less effective.

To address these issues, we have developed a stock prediction framework that addresses these issues by (1) constructing features that capture both the sentiment and online status of posters in social media, and (2) processing features using cumulative sum (CUSUM) (Page, 1954) and exponential weighted moving average (EWMA) (Roberts, 1959) in order to enhance weak signals and improve prediction performance. We then conducted experiments of the stock movement prediction and the simulated trading to test the effectiveness of our proposed method. We evaluated our frameworks based on 23 stocks in the New York Stock Exchange (NYSE) and NASDAQ that had high message volumes on Yahoo! Inc. message boards.

Our study makes several contributions. First, we develop novel enhanced features involving both sentiment and online social relations for stock predictions. Second, we report empirical evidence indicating that our design, especially CUSUM, helps to predict stock returns better than other settings without such enhancing approaches. Third, we test our proposed approach and find that in a simulated trading that considers transaction costs using the high-frequency bid and ask prices, it outperforms other baseline approaches.

The rest of this study is structured as follows. First, in Section 2, we briefly review the related literature. Then, in Section 3, we discuss the research questions and detail our stock prediction framework. In Section 4, we present the experimental designs and detail the experimental results of our proposed models compared to a baseline. Finally, Section 5 presents our conclusion.

2. LITERATURE REVIEW

In this section, we review relevant studies under the following four categories: (1) the relationship between public information and stock markets, (2) the feature representations for text data, (3) online social relations on social media, and (4) statistical surveillance approaches.

2.1 Public information and stock markets

Recent studies have shown that public information including firm-specific news (Geva & Zahavi, 2014; Hagenau et al., 2013) and social media (microblogs and discussion boards) (Das & Chen, 2007; Smailović et al., 2014) can influence stock prices, volumes, and volatility. Firm-specific news promptly reports information about fundamentals and recent activities of a firm that enters stock market participants' information sets (Geva & Zahavi, 2014; Groth et al., 2014; Q. Li et al., 2014). Several studies have reported that firm-specific news can predict stock price movements after a news article has been released (Q. Li et al., 2014; Schumaker & Chen, 2009; Schumaker et al., 2012).

Unlike firm-specific news that is compiled by professional writers, social media content reflects the opinions of individual users and is usually generated through peer interactions. Recent studies have documented how message volumes and sentiments, represented as daily time series, are correlated to the values of the stock index (Bollen et al., 2011; Das & Chen, 2007), trading volumes and volatility (Das & Chen, 2007), and firm-specific closing price changes (Smailović et al., 2014).

2.2 Feature representations for text data

A stock prediction model needs to adopt a suitable feature representation for news and social media content. Such a suitable feature representation can improve prediction performance by capturing relevant information and reduce potential downsides, such as the sparsity issue caused by high dimensions of features (Forman, 2003). Popular feature representations include sentiment, latent topics, proper nouns, and N-grams.

Sentiment-based feature representations have been popular in previous research (Abrahams et al., 2012; Choi et al., 2009), and investment decisions can be influenced by public sentiment in social media and news articles (Bollen et al., 2011; Hagenau et al., 2013). Words such as "bull" and "bear" can be used to identify the positive and negative sentiment in a sentence, paragraph, or document. One approach is to count words from a sentiment dictionary, such as *Harvard-IV-4* (Stone et al., 1966) or Loughran and McDonald's *Financial Sentiment Dictionary* (Bodnaruk et al., 2015). Another technique uses a machine learning approach to automatically classify words into positive and negative sentiments based on the direction of the stock price movement (Q. Li et al., 2014).

Topic-based representation, such as latent topics discovered by latent Dirichlet allocation (LDA) (Blei et al., 2003), groups words into topics and adopts these topics as their predictive features (Jin et al., 2013). Latent topics can be regarded as a variety of specific events and trends that are captured from financial news and social media content. Jin et al. (2013) showed that the financial market is sensitive to several types of events in the news. Topic-based features can forecast the movement of stock prices (Lavrenko et al., 2000) and foreign currency markets (Jin et al., 2013).

Schumaker and Chen (2009) employed another types of representation, involving noun phrases, named entities, and proper nouns, for their predictive models. Noun phrases selects only the nouns and noun phrases of an article; named entities has nouns or noun phrases selected in pre-identified categories tagged using a semantic lexicon; and proper nouns involves particular nouns that refer to a unique entity, such as the names of people, companies, locations, or organizations. Schumaker and Chen (2009) found a proper nouns approach to be the most effective of these three for stock price predictions.

Finally, another feature representation focuses on individual words or word combinations used in articles. One of the most commonly adopted representations of this type is bag-of-words, which represents the occurrence of individual words (Groth & Muntermann, 2011). N-gram, an extension of bag-of-words, combines a sequence of N words. While N could be any integer, 2-Gram is the most commonly used type (Hagenau et al., 2013).

2.3 Online social relations

Online social relations can affect customers' purchase decision making (Y.-M. Li et al., 2013) and influence the future popularity of online items, such as online videos clips (Susarla et al., 2012). Forming social ties with high-popularity users leads the user to have a more similar preference to high-popularity users (Zeng & Wei, 2013). Users with higher social status also have a higher influence than other users in a network (Lee et al., 2011). Empirical evidence suggests that a large part of user decision making is dependent on the decisions of those with higher status and popularity (Susarla et al., 2012). While few studies have investigated how online social relations influence investment decision, we hypothesize that the opinions from posters with higher online social status may be more influential to others' investment decisions and it is more important for stock prediction. Among the popular measures that have been used to capture the characteristics of social networks are outdegree (Hanneman & Riddle, 2005), indegree (Hanneman & Riddle, 2005), pagerank (Brin & Page, 1998), authority and hub (Kleinberg, 1999), betweenness centrality (Brandes, 2001) and closeness centrality (Hanneman & Riddle, 2005).

Outdegree gauges the centrality of a node by counting the number of links pointing out from it, and indegree measures the prestige of a node by the number of links connecting to it in a network (Hanneman & Riddle, 2005). Pagerank measures the popularity of a node by recursively computing the prestige scores in a network (Brin & Page, 1998). Specifically,

$$P(n) = \frac{(1-d)}{N} + d\sum_{x\in M(n)} \frac{P(x)}{\mathrm{OD}(x)}, \quad (1)$$

where $P(n)$ is the pagerank score of node n, d $(0 \leq d \leq 1)$ is a tuning parameter, N is the number of total nodes, $\mathrm{OD}(x)$ is the number of outdegrees of node x, and $\mathrm{M}(n)$ is the set of nodes that link to node n.

Authority $(A(n))$ and hub $(H(n))$ both measure the authority of a node in a network (Kleinberg, 1999). The equation used for these is

$$A(n) = \sum_{x \in M(n)} H(x)\,, and\ H(n) = \sum_{x \in M(n)} A(x), \tag{2}$$

where $A(n)$ is the authority score, $H(n)$ is the hub score of node n, and $M(n)$ is the set of nodes reacting to node n.

Betweenness centrality computes the shortest paths of node n among all the nodes in a network (Brandes, 2001). This employs

$$B(n) = \sum_{i<j, i \neq n, j \neq n} \frac{path_{ij}(n)}{path_{ij}}, \tag{3}$$

where $B(n)$ is the betweenness centrality score of node n, $path_{ij}$ is the number of shortest paths between node i and node j, and $path_{ij}(n)$ is the number of paths between node i and node j that pass through node n.

Closeness centrality measures the distance of node n to all others in a network (Hanneman & Riddle, 2005). This uses

$$C(n) = \left[\sum_{x \neq n}^{N} d(n,x)\right]^{-1}, \tag{4}$$

where $C(n)$ is the closeness centrality score of node n, N is the number of total nodes, and $d(n,x)$ is the shortest path between node n and node x.

Generally, these online metrics compute the connectivity, importance and centrality of a node with their neighbors in a social network (Ma et al., 2009).

2.4 Statistical surveillance approaches

Statistical surveillance approaches are used extensively in monitoring process variability (Montgomery, 2005). A common scenario is to supervise the quality of a process or product by monitoring the mean or variance of sequentially selected observations (Koutras et al., 2007). There are two major scenarios for the applications of statistical surveillance approaches. The first scenario is to detect large shifts, using approaches such as the Shewhart control chart. The second scenario is to detect small shifts, using approaches such as cumulative sum (CUSUM) (Page, 1954) and exponential weighted moving average (EWMA) (Roberts, 1959).

A Shewhart control chart considers only the last inspected sample and ignore the whole sequence of observations that is less sensitive for small shift detection (Montgomery, 2005). In contrast, CUSUM is concerned with the change of entire sequent observations and accumulates the deviations, up and down, of samples from a target mean (Montgomery, 2005). EWMA also takes into account cumulative changes, calculating weighted averages for all past observations over time (Montgomery, 2005). Apart from applying them to the process control, CUSUM and EWMA are also efficient for detecting changes of specific word occurrences from textual streams on social media to capture emerging trends (Denecke et al., 2013; Schubert et al., 2014). In this study, we adopt both the CUSUM and EWMA approaches to capture emerging trends based on our text sentiment and online social relation features constructed from social media.

3. RESEARCH FRAMEWORK

3.1 Research questions

Previous studies in stock predictions have mostly focused on how text features, such as sentiment and volume, influence stock markets. Few studies have investigated the value of online social relations in predicting stock movements. Our goal here was to develop a framework that incorporate online social relations in stock movement predictions. We asked the following research questions:

- How do we incorporate online social status into predictive models?
- Can statistical surveillance approaches improve prediction ability in this context?
- Does the newly proposed approach have better profitability than baselines using traditional sentiment features?

3.2 System design

To answer our research questions, we developed a novel system that extracts message volume (V_SM), sentiments (S_SM), and online social relations (GRAPH) from a social media platform. We also incorporated sentiments (S_News) from firm-specific news and historical market data based on the guidance of previous studies (Hagenau et al., 2013; Schumaker et al., 2012). In order to further enhance the ability to capture emerging trends, we adopted the statistical surveillance approaches including CUSUM and EWMA to further enhance signals embedded in these features. We constructed predictive models using support vector regression (SVR) (Vapnik, 1995) to evaluate the utility of our design. Finally, we conducted simulated trading to investigate the potential profitability of the proposed design framework. Figure 1 depicts this system design.

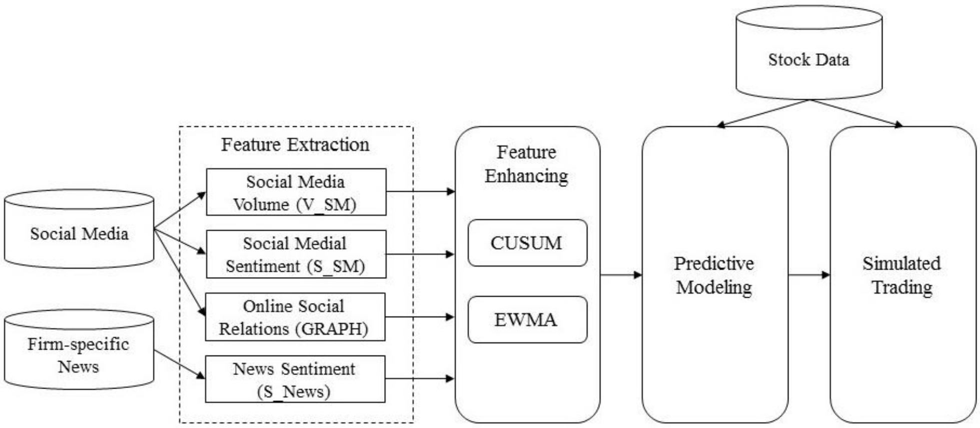

Figure 1. *System design.*

3.2.1 Feature extraction

To facilitate subsequent high frequency stock price prediction and simulated trading, we adopted a short time window of length *l* during trading hours and indexed these time intervals by $t \in \{0, 1, 2, \cdots, T\}$. We set *l* to 20 minutes, following Schumaker and Chen (2009). We captured sentiment and volume from content derived from social media and news articles (S_SM, V_SM, and S_News) during time interval t to predict the stock return in the following 20-minute time interval. However, since extracting online social relations using a short window of 20 minutes would produce extremely sparse social networks that would be of limited use for our subsequent social enhancing approach, we extracted online GRAPH features based on social media content over the previous six months. Here, we detail our V_SM and S_SM, GRAPH, and S_News features.

Social media volume (V_SM) and sentiment (S_SM)

V_SM is the number of messages on a social media regarding a stock. S_SM extracts the sentiment of the text content on social media. S_SM contains three components: positive message volumes at time interval t (PS(t)), negative message volume at time interval t (NS(t)), and the difference between positive and negative messages (DS(t) = PS(t)-NS(t)). We obtained these counts through a sentiment classifier trained on labeled messages from the focal social media platform. We constructed our sentiment classifier using LIBSVM (Chang & Lin, 2011), a popular support vector machine (SVM) implementation. Since a portion of the messages in our focal social media (Yahoo Inc. Message Board) contained sentiment labels, we adopted these labeled messages as our training instances to construct the sentiment classifier and applied the learned classifiers on all messages to obtain sentiment counts in each time interval.

Online social relations (GRAPH)

GRAPH captures the online social relations of posters on social media. We assume a directed link from user A to user B if user A reacted to a message from user B. GRAPH contains seven metrics: outdegree, indegree, pagerank, authority, hub, betweenness, and closeness. Outdegree is the number of users to whom a user u reacts, and indegree is the number of users who react to the focal user; pagerank is the pagerank score of a user and can measure the user's popularity (Brin & Page, 1998); authority and hub both represent the authority of a user (Kleinberg, 1999); betweenness measures the importance of a user in connecting other users (Brandes, 2001); while closeness measures how easily a user can reach other users, and can represent the centrality of the user in the network (Hanneman & Riddle, 2005).

To compute the GRAPH metrics at time interval t, we extracted all users who posted or replied to messages in that time interval and computed the summation of each metric over these users. The social network was constructed based on the reply activities during the previous six months. We adopted a network analysis package, igraph, to facilitate the computation of network metrics.

News sentiment (S_News)

We adopted two approaches to extract news sentiment. The first approach was OpinionFinder (Wilson et al., 2005), a popular sentiment tool. OpinionFinder conducts token-level sentiment labeling based on the context of a sentence. We counted the positive words (PS_OP), negative words (NS_OP), and the difference between positive and negative words (DS_OP) in news articles released during a time interval. The second approach was based on the financial sentiment dictionary (Bodnaruk et al., 2015). We recorded the positive words (PS_FD), negative words (NS_FD), and the difference between the positive and negative word counts (DS_FD) for news articles released in the time interval.

Feature normalization

A previous study (Antweiler & Frank, 2004) suggested that there were strong intra-day cyclical patterns in social media platforms. Our preliminary investigation also suggested that our research testbed had a similar intraday patterns. To address this issue, we adopted a feature normalization approach to correct for the intraday cyclical dynamics.

For each metric in V_SM, S_SM, GRAPH, and S_News, we normalized the feature value by computing the difference between its input value and its six-month moving average. We computed the six-month moving average using only the values of the same time-of-day during the previous six months. For example, to normalize outdegree for time interval t corresponding to the period 10:00 am to 10:20 am of that day, the all outdegree values corresponding to the 10:00 am to 10:20 am period over the previous six months were used to compute the moving average.

3.2.2 Feature enhancing via statistical surveillance approaches

In order to enhance the social media signals, we adopted CUSUM and EWMA to process the normalized feature values. The basic idea here was to accumulate small systematic trends so that these changes could be captured early. The key parameter for our procedure was historical length, W. For a specific W, we computed the CUSUM score of a feature starting from $t - W$ to t.

Specifically, the upper CUSUM value was $C_{it}^{W+} = \max\left[0, x_{it} - (u_i + K) + C_{it-1}^{W+}\right]$, where x_{it} is the normalized value of feature i at time interval t, u_i is the mean of the normalized feature value in the past six months, and K is the one-half of the standard deviation of the normalized feature value in the past six months. We set C_{it-W}^{W+} to 0, since $t - W$ was the starting period. Similarly, we computed the lower CUSUM value $C_{it}^{W-} = \min\left[0, x_{it} - (u_i - K) + C_{it-1}^{W-}\right]$, and used $C_{it-W}^{W-} = 0$. We considered $W = 1, 2, 3, \dots, 10$ in our study.

We computed EWMA score using

$$z_{it}^{W} = \lambda x_{it} + (1 - \lambda) z_{it-1}^{W}, \tag{5}$$

where z_{it}^{W} is an EWMA value of feature i at time interval t using a historical window W. The parameter λ should take a value between 0 and 1. We set λ to 0.8, based on parameter tuning using training datasets.

3.2.3 Predictive tasks and simulated trading

We constructed a high frequency stock return prediction model that predicts the following 20-minute (period $t + 1$) returns based on currently (period t) available information. We defined stock return as $r_{t+1} = \frac{bid_{t+1} - ask_t}{ask_t}$, where bid_{t+1} is the bid price at period $t + 1$ and ask_t is the ask price at period t. This definition was based on a real life scenario in which an investor needs to pay the ask price to buy stock and takes the bid price to sell it. This definition incorporates a significant portion of transaction costs into stock returns and is less susceptible to the unrealistic optimism where high frequency returns are defined using transaction prices only.

We incorporated all the enhanced features—V_SM, S_SM, GRAPH and S_News—together with the historical market data in our model. After CUSUM and EWMA processing, each metric was expanded to ten scores of upper CUSUM, lower CUSUM, and EWMA, corresponding to ten different W values. The historical market data comprised the trading volume, bid price, ask price, and stock return during period t.

At the beginning of each month, we trained our model using the data in the past 12 months and used this to predict high frequency returns in that month. We tuned SVR parameters using ten-fold cross validation on the training dataset. The model setting is discussed in detail below (Section 4.2).

In order to evaluate the potential profitability of the proposed design framework, we also conducted a simulated trading process. We utilized the trained predictive model to generate trading signals. If the predicted value was greater than the upper threshold, then the system bought the stock immediately and disposed of the holding in the next 20-min time interval. If the predictive price was lower than the lower threshold, then the system short sold the stock and disposed of the holding in 20 minutes. We tuned the threshold using the previous month's training data.

4. EXPERIMENTAL EVALUATION

4.1 Data set

We constructed three datasets that combined stock message boards, firm-specific news articles, and stock market data for our experiments. Our testbed spanned a period of 100 months (2,094 trading days), from January 1, 2000 to April 30, 2008. Following previous studies (Schumaker & Chen, 2009; Schumaker et al., 2012), the daily experimental periods for predicting stock returns was restricted to between 10:00 am and 3:40 pm in a trading day, in order to avoid opening and closing quotes. In this work, we targeted 23 companies listed on New York Stock Exchange (NYSE) and NASDAQ (see Table 1). The companies were selected because of their high message volumes on stock message boards. This selection was consistent with the previous work (Das & Chen, 2007). For the selected period, we gathered 20,846,478 messages from Yahoo! Inc. Message Boards and 100,615 news articles from Reuters news that were relevant to the target companies. The collected messages contained both users' posts and other users' comments. We also obtained the transaction records of stock markets including transaction prices, bid prices, ask prices, and trading volumes from the New York Stock Exchange Trades and Quotes (TAQ) database. We calculated the target stock return by using bid and ask prices for prediction, which thus took into account the transaction cost for each trading. Table 2 summarizes our dataset.

Stock exchange	Stock ticker
NASDAQ	AAPL, AMZN, CSCO, EBAY, ET, INTC, JAVA, JDSU, MSFT, QCOM, RIMM, RMBS, SIRI, YHOO
NYSE	ALU, CPN, GE, ELN, NEM, NT, TWX, TYC, WEL

Table 1. The target 23 companies.

Dataset	Data source	Feature type	Feature name
Stock message boards	Yahoo! Inc. Message Boards	V_SM	Message volume
		S_SM	PS, NS and DS
		GRAPH	Outdegree, indegree, pagerank, authority, hub, betweenness, and closeness
Firm-specific news	Reuters news	S_News	PS_OP, NS_OP, DS_OP, PS_FD, NS_FD and DS_FD
Stock market data	New York Stock Exchange Trades and Quotes (TAQ) database	Historical market data	Trading volume, bid price, ask price, and return

Table 2. Three experimental datasets.

4.2 Design

To investigate our research questions, we developed two feature-enhanced models, namely, SE_CUSUM and SE_EWMA, to predict high frequency returns. SE_CUSUM contained historical market data and V_SM, S_SM, GRAPH, and S_News features enhanced by CUSUM. SE_EWMA contained historical market data and EWMA-enhanced V_SM, S_SM, GRAPH, and S_News features.

In order to evaluate the individual effect of the sentiment and the online social relation in stock prediction, we divided our enhanced features into two types of models. The first incorporated historical market data with S_SM, and S_News features enhanced by CUSUM and EWMA; these are referred to here as S_CUSUM and S_EWMA. The second combined historical market data with GRAPH features enhanced by CUSUM and EWMA (G_CUSUM and G_EWMA).

We compared our proposed models with six baseline models. The first baseline model was All_Metrics, which contained historical market data and features in V_SM, S_SM, GRAPH, and S_News. These features were not processed by the statistical surveillance approaches. The second baseline was Historical_Market, which included only historical market data. The remaining baselines were random guessing (RG), guessing positive (GP), guessing negative (GN), and guessing zero (HZ). RG predicted the direction (up or down) randomly; GP always predicted a positive return; GN always predicted a negative return; and HZ always predicted the following period return to be zero.

4.3 Performance evaluation

We adopted four evaluating measures to gauge our predictive performance, including root mean square error (RMSE), mean absolute error (MAE), directional accuracy, and balanced error rate (BER). RMSE was the square root of the mean squared difference between predicted and true stock return, while MAE was the average absolute difference between the predicted and actual returns. The predicted result sometimes came close to the target value but in the wrong direction. Directional accuracy computed the proportion of currently predicted return directions compared with the true directions, and BER (Chen & Lin, 2006) measured the balanced error rate of both positive and negative stock returns:

$$\text{BER} = \frac{1}{2}\left(\frac{\#\text{ positive instances predicted wrong}}{\#\text{ positive instances}} + \frac{\#\text{ negative instances predicted wrong}}{\#\text{ negative instances}}\right). \tag{6}$$

To evaluate trading performance, we simulated a trading process with the proposed models and compared this to the baseline methods. We endowed each model an initial investment budget of 1,000 USD, which was equally assigned to every company in our testbed. The investment period ran from January 1, 2001 to April 30, 2008, a period of 1,840 trading days. Each model predicted the stock return for the 20-minutes time interval between the hours of 10:00 am and 3:40 pm, and buy or short sell the stock based on the predicted returns. We compare different models according to their trading profits.

4.4 Results

Table 3 reports the prediction performance of our proposed models and baseline models. First, both SE_CUSUM and SE_EWMA performed better than did the baselines. SE_CUSUM had the best directional accuracy, BER, RMSE, and MAE compared to the SE_EWMA and baseline models. Second, SE_CUSUM outperformed S_CUSUM and G_CUSUM in the four evaluating measures, suggesting that considering both sentiment and online social relations for stock predictions is more effective than directly adopting individual types of features. SE_EWMA also outperformed S_EWMA and G_EWMA (further detailing of that here is precluded for reasons of space). Third, RG had the worst directional accuracy (0.4901), while GP was slightly better (0.4915). The directional accuracy of GN (0.5085) was better than GP, suggesting that our stock return definition using bid and ask prices may tend to have negative returns. Another interesting observation was that HZ (predicting zero return) performed better than S_CUSUM, G_CUSUM, All_Metrics, and Historical_Market. Since the random walk hypothesis implies predicting zero return is the best strategy, the relatively better performance of HZ suggests that the random walk hypothesis may be useful when the goal is not to conduct high frequency trading.

We performed a two-tailed, paired *t*−test to investigate the statistical significance of performance differences. As reported in Table 3, the results for this showed that SE_CUSUM was significantly better than SE_EWMA, S_CUSUM, G_CUSUM, and all baselines, at a 5% significance level. SE_EWMA was also significantly better than S_EWMA, G_EWMA, and all baselines at a 5% significance level. (Again, we do not report the details of the comparison of SE_EWMA with other models due to the space limitations.) These results suggest that CUSUM is an effective approach to enhance signals embedded within social media. EWMA can also improve performance but is less effective than CUSUM.

Method	Directional accuracy	BER	RMSE	MAE
SE_CUSUM	**0.5720**	**0.4423**	**0.011293**	**0.009622**
SE_EWMA	0.5689**	0.4455**	0.011567*	0.009974*
S_CUSUM	0.5563***	0.4569***	0.012118**	0.010443**
G_CUSUM	0.5504***	0.4616***	0.012481***	0.010716***
All_Metrics	0.5643***	0.4487***	0.011856**	0.010146**
Historical_Market	0.5316***	0.4752***	0.013896***	0.012141***
RG	0.4901***	0.4994***	-	-
HZ	-	-	0.011765**	0.010083*
GP	0.4915***	0.5***	-	-
GN	0.5085***	0.5***	-	-

* Indicate the difference between SE_CUSUM and a method is statistically significant at p <0.05.
** Indicate the difference between SE_CUSUM and a method is statistically significant at p <0.01.
*** Indicate the difference between SE_CUSUM and a method is statistically significant at p <0.001.

Table 3. Predicting results of feature-enhanced modes versus baseline models.

Table 4 shows the results for the simulated trading experiments. In these experiments, we selected the two feature-enhanced models SE_CUSUM and SE_EWMA and compared these to the baseline models. The baseline models included All_Metrics and Historical_Market only. Since HZ is unable to generate trading signals. RG, GP, and GN cannot compute RMSE and MAE, they use directional predictions only. Consistent with the result of the prediction performance, SE_CUSUM had the best end-of-period portfolio value (1,776.1 USD), followed by SE_EWMA (1495.0 USD). Historical_Market had a portfolio value of 835.4, which corresponds to a return of -0.1645. The results clearly suggest that SE_CUSUM had a better simulated trading returns than other approaches.

Figure 2 shows the monthly trading results of these four methods over the investment period calculating the monthly profits that accumulated earnings over previous trading days. The cumulative portfolio values of SE_CUSUM had a clear increment over our investment period; SE_CUSUM also showed an obviously better trend than other methods.

Method	Portfolio values	Return
SE_CUSUM	1766.1	0.7661
SE_EWMA	1495.0	0.4949
All_Metrics	1180.7	0.1806
Historical_Market	835.4	-0.1645
Initial investment budget USD 1,000 for each method and equally assigned to each company.		

Table 4. The simulated trading result.

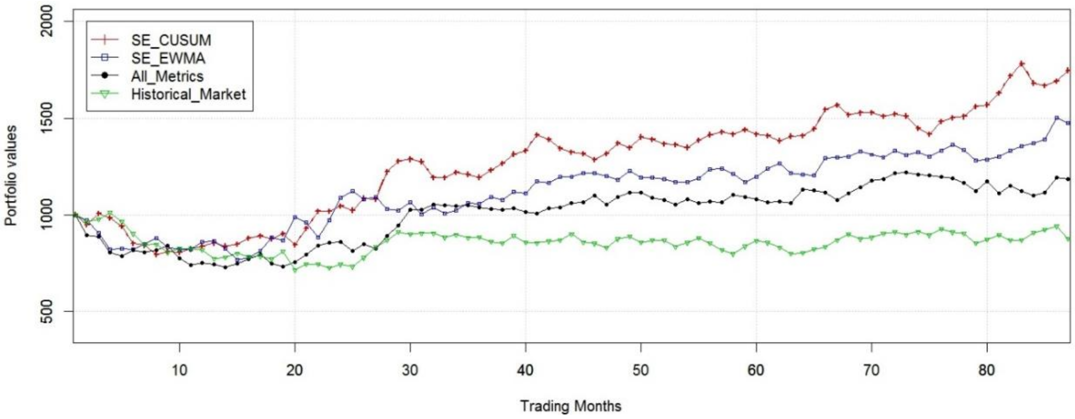

Figure 2. Simulated trading results in the 88 months of the investment period.

5. CONCLUSION

We have made several contributions in this paper. First, we have developed features that captured the online social status of users engaged in content generation. These features were further processed by statistical surveillance approaches to boost their ability to capture emerging trends. Second, SE_CUSUM, the feature-enhancing design based on CUSUM, was found to outperform baseline models in predicting stock returns in the following 20 minutes. The evidence suggests that our feature-enhancing design, especially CUSUM, along with the online social status features are more effective for stock predictions. Third, in the simulated trading experiment, SE_CUSUM was shown to have the best end-of-period portfolio value, achieving a positive return of 0.7661 and providing promising evidence that the proposed approach may profit in high frequency stock trading even when bid-ask spreads are incorporated in transaction costs.

References

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems, 54*(1), 87-97.

Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance, 59*(3), 1259-1294.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine learning research, 3*, 993-1022.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis, 50*(04), 623-646.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology, 25*(2), 163-177.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 27.

Chen, Y.-W., & Lin, C.-J. (2006). Combining SVMs with various feature selection strategies: *Feature extraction* (pp. 315-324): Springer.

Choi, Y., Kim, Y., & Myaeng, S.-H. (2009). *Domain-specific sentiment analysis using contextual feature generation.* Paper presented at the Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science, 53*(9), 1375-1388.

Denecke, K., Krieck, M., Otrusina, L., Smrz, P., Dolog, P., Nejdl, W., & Velasco, E. (2013). How to exploit twitter for public health monitoring. *Methods Inf Med, 52*(4), 326-339.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research, 3*, 1289-1305.

Geva, T., & Zahavi, J. (2014). Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision Support Systems, 57*, 212-223.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems, 50*(4), 680-691.

Groth, S. S., Siering, M., & Gomber, P. (2014). How to enable automated trading engines to cope with news-related liquidity shocks? Extracting signals from unstructured data. *Decision Support Systems, 62*, 32-42.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685-697.

Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods: University of California Riverside.

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., & Ramakrishnan, N. (2013). *Forex-foreteller: Currency trend modeling using news articles.* Paper presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604-632.

Koutras, M. V., Bersimis, S., & Maravelakis, P. E. (2007). Statistical Process Control using Shewhart Control Charts with Supplementary Runs Rules. *Methodology and Computing in Applied Probability, 9*(2), 207-224.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000). *Mining of concurrent text and time series.* Paper presented at the KDD-2000 Workshop on Text Mining.

Lee, M. K. O., Shi, N., Cheung, C. M. K., Lim, K. H., & Sia, C. L. (2011). Consumer's decision to shop online: The moderating role of positive informational social influence. *Information & Management, 48*(6), 185-191.

Li, Q., Wang, T., Gong, Q., Chen, Y., Lin, Z., & Song, S.-k. (2014). Media-aware quantitative trading based on public Web information. *Decision Support Systems, 61*, 93-105.

Li, Y.-M., Wu, C.-T., & Lai, C.-Y. (2013). A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship. *Decision Support Systems, 55*(3), 740-752.

Ma, Z., Sheng, O. R. L., & Pant, G. (2009). Discovering company revenue relations from news: A network approach. *Decision Support Systems, 47*(4), 408-414.

Montgomery, D. C. (2005). *Introduction to statistical quality control* (5th ed.): John Wiley.

Page, E. (1954). Continuous inspection schemes. *Biometrika, 41*(1/2), 100-115.

Roberts, S. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics, 1*(3), 239-250.

Schubert, E., Weiler, M., & Kriegel, H.-P. (2014). *SigniTrend: scalable detection of emerging topics in textual streams by hashed significance thresholds.* Paper presented at the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems, 27*(2), 1-19.

Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458-464.

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences, 285*, 181-203.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A Computer*

Approach to Content Analysis: M.I.T. studies in comparative politics. M.I.T. Press, Cambridge,MA.

Susarla, A., Oh, J.-H., & Tan, Y. (2012). Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube. *Information Systems Research, 23*(1), 23-41.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*: Springer.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). *OpinionFinder: A system for subjectivity analysis.* Paper presented at the Human Language Technology Conference.

Zeng, X., & Wei, L. (2013). Social Ties and User Content Generation: Evidence from Flickr. *Information Systems Research, 24*(1), 71-87.